J. Appl. Environ. Biol. Sci., 6(3S)114-120, 2016

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

Line and Ligature Segmentation in Printed Urdu Document Images

Israr Ud Din, Zumra Malik, Imran Siddiqi and Shehzad Khalid

Bahria University, Islamabad, Pakistan

Received: January7, 2016
Accepted: March 2, 2016

ABSTRACT

This paper presents a technique for segmentation of printed Urdu text images into lines and ligatures, a key pre-processing step in Urdu Optical Character Recognition (OCR) systems. Unlike classical projection profile based line segmentation methods, the proposed scheme successfully segments overlapping and touching lines. Once the lines are segmented, ligatures are extracted from each text line by associating the secondary ligatures with their respective primary ligatures. The system evaluated on 30 printed Urdu documents with 310 text lines and 7,364 ligatures realized promising results on line and ligature segmentation.
KEYWORDS: Urdu OCR, pre-processing, line segmentation, ligature segmentation

1. INTRODUCTION

Optical Character Recognition (OCR) is one of the most researched and classical pattern recognition problems which includes conversion of an image of handwritten or printed text into machine readable form [3, 9]. Images with textual content include electronic fax files, scanned documents or documents/text captured through a camera. Text in the form of an image is of less value as it is neither editable nor searchable. One solution could be the manual transcription of huge collections of digitized documents which naturally is a tedious task requiring massive human effort. Consequently, automated recognition engines were researched and developed to convert images into text. These OCR systems find applications in a wide range of areas including sorting of postal mails, automatic reading of filled forms, digital libraries and license plate recognition etc. A typical OCR system mainly comprises three key steps including pre-processing, recognition and post-processing [4].

OCR is considered a mature research area for text in a number of languages around the globe. Today, commercial OCR systems reporting near to 100% recognition rates are available for languages based on the Latin alphabet (English, German, French etc.) Arabic and Chinese OCR systems are also very mature realizing high recognition rates. Despite the significant research attention this problem as attracted, OCR systems for a large number of scripts are either non-existent or are in very early stages. One such language is Urdu and makes the subject of our present study.

Urdu, a major language of the Indian sub-continent is spoken by millions of individuals around the globe. Urdu comprises 39 basic characters including 28 characters from Arabic. Urdu is mostly written in the Nastaliq writing style as opposed to Arabic which mostly follows the Naskh style of writing [10]. Both Urdu and Arabic are context sensitive languages where each character may take up to 4 different shapes depending upon its position (isolated, initial, middle, and end) within a word. The basic units of recognition in Urdu are ligatures where each ligature comprises one or more characters joined together. A set of sample Urdu ligatures is illustrated in Figure 1. Ligatures are further categorized into primary and secondary ligatures. Primary ligatures represent the main body while secondary ligatures correspond to the dots or diacritic marks.

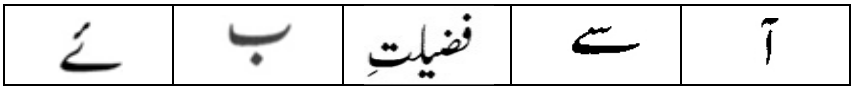


Figure 1. Examples of Urdu ligatures

The highly cursive nature of Urdu text, context sensitivity of characters, frequent presence of dots and diacritic marks, non-uniform inter and intra word spacing and overlap of ligatures across multiple lines make segmentation of Urdu text a challenging problem. An effective segmentation of document into lines and subsequently into ligatures is vital for recognition. Majority of the existing work relies on simple horizontal

* **Corresponding Author:** Israr Ud Din , Bahria University, Islamabad, Pakistan
israr.uddin@yahoo.com,

projection method for line segmentation [4, 6, 7, 8]. These methods, however, cannot handle situations where the words from different lines overlap. This study presents a hybrid method based on horizontal and vertical projections to segment touching and overlapping text lines in digitized Urdu documents. A scheme for segmentation of ligatures and association of dots/diacritics with the main ligature body is also proposed. The proposed segmentation evaluated on a standard database of scanned Urdu books realized promising results on line and ligature segmentation.

2. RELATED WORK

Segmentation of text into lines and ligatures is the key pre-processing step in any recognition system. A number of studies on Urdu OCR, therefore, employ different techniques to carry out this segmentation. Among well-known segmentation methods, Pal and Sarkar [7] present a line segmentation method based on projection profiles which achieves a correct segmentation of 98%. The technique, however, fails in cases where ligatures of adjacent text lines overlap.

Malik et al. [6] employ left-to-right horizontal image scanning for text-line detection. The technique searches for a row with minimum text pixel count and if a secondary ligature is encountered, it is assigned to the nearest ligature among the candidate text lines. The authors, however, do not mention the segmentation accuracy of their proposed system. In a similar study by Javed and Hussain [4], projection features are used to segment text lines while dots/diacritics are associated to the respective primary ligatures using vertical distances. The authors report 100% line segmentation accuracy but the method is not evaluated on text images with touching characters across multiple lines.

In another study [8], the authors binarize gray scale document images and extract text-lines using horizontal projections. White spaces are removed using a set of heuristics and the dots are associated with their corresponding primary ligatures. An accuracy of 97% is reported by this system. Bukhari et al. [2] present a comprehensive system for document layout analysis including segmentation of text and non-text regions, text-line segmentation and determination of reading order. The system evaluated on 25 Arabic documents for text/non-text separation realized an accuracy of 99%. For text line segmentation, segmentation accuracies of 96% and 92% are reported on 25 Arabic and 20 Urdu documents respectively.

Adiguzel et al. [1] propose a mixture of connected component and projection information for segmenting lines in historical Ottoman documents. The proposed technique evaluated on a custom data set consisting of both printed and handwritten books reports an overall line-segmentation accuracy of 95%. An interesting aspect of this study is that the segmentation results on printed and handwritten documents are of the same order and are invariant to writing style and skewness. In a relatively recent study [5], a technique for segmentation of Urdu text into lines and ligatures is presented. The method is robust to handle overlapped ligatures and successfully segments touching lines as well. The system is evaluated on 448 Urdu text lines and realized an accuracy of 99%.

After having discussed some notable contributions towards segmentation of Urdu text, we present the proposed segmentation scheme in the following section.

3. Proposed Segmentation Methodology

We carry out segmentation at two levels, line and ligature. The given document image is first binarized using global thresholding and lines are segmented. From each segmented line, the ligatures are then extracted. Each of these modules is discussed in detail in the following.

3.1 Line Segmentation

Typically, text lines are extracted by computing the horizontal projection profile of an image (counting the number of text pixels in each row of the image). The local peaks (row with maximum number of text pixels) and local valleys (row with minimum number of text pixels) of the projection profile are then analyzed to find the segmentation point [4,7]. This scheme works fairly well for text in many scripts. However, in case of Urdu text, the presence of gaps or white spaces between primary and secondary ligatures in a text-line results in multiple peaks and valleys as illustrated in Figure 2.

Figure 2: Multiple peaks and valleys in same line of text due to dots and diacritics

Carrying out segmentation at valleys of the horizontal projection results in over segmentation an example being shown in Figure 3. Not only such situations lead to incorrect segmentation of lines but the association of secondary ligatures (dots and diacritics) to their respective ligatures also becomes challenging.

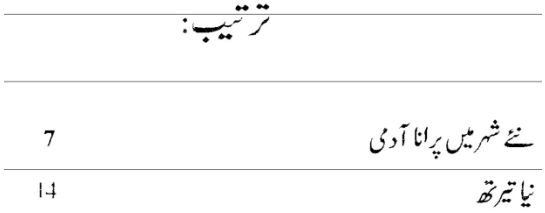


Figure 3: Over-segmentation of first line due to presence of dots

To overcome the aforementioned issues, we propose a segmentation technique based on a combination of projection profiles and a set of heuristics [5]. The segmentation of zones using estimated row height proposed in [5], in some cases, results in under segmentation especially if the image has noisy components. We, therefore, first perform a crude segmentation of the binarized document image using the traditional horizontal projection method. The height of each resulting zone is then computed as in [5] and the median zone height in the document is determined.

In order to avoid the problems with assignments of dots and diacritics and over and under segmentation, we dilate the document image with a square structuring element to join neighboring ligatures and words as well as the secondary ligatures with the primary components. An example of an original and dilated image is illustrated in Figure 4. Median zone height is then used as threshold for finding local peaks in dilated version of the document image which in turn is used for finding the valley index (that is the probable text-line boundary) between two consecutive local peaks in the image. Zones deviating significantly from the median zone height are identified as under segmented lines.

Figure 4: Original image and its dilated version

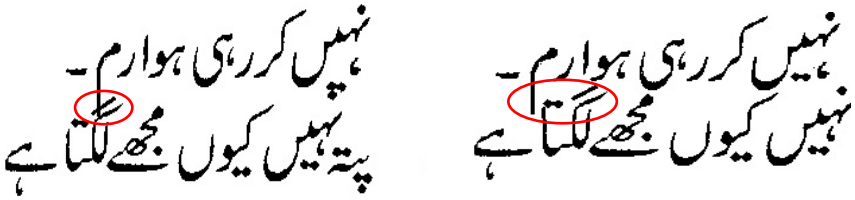
The segmentation carried out after dilating the image works reasonably good for documents without overlapped lines. However, for documents with a larger number of overlapped text lines, using the median zone height as threshold for identifying the largest peak in the projection profile may lead to under

segmentation. An example of this scenario is illustrated in Figure 5 where most of the lines in the document are overlapping and the zone height computed from projection profile is misleading causing an under segmentation (Figure 5).


عامر: چپ چپ تھکی تھکی.....واپس چلیں۔
ارم: نہیں۔
عامر: نہیں ہوا ہے ارم۔
ارم: بھئی کچھ نہیں۔
عامر: پنڈ والوں کا کھانا کچھ کمزور سا ہوتا ہے۔تندور کی روٹی اور کچی پکی مرغی۔تم Mind نہیں کر رہی ہو ارم۔
ارم: پتہ نہیں کیوں مجھے لگتا ہے عامر جیسے میں آخری بار اس جنت کو دیکھ رہی ہوں۔
عامر: وہم نہیں کرتے اندر چلو۔
ارم: فارم میں گھستے ہی اگر واجد نے بندوق نہ چلا دی ہوتی ، پتہ نہیں فاختہ مرگئی یا گھر پہنچ گئی۔شغالے کے کھیت میں گرتی تو میں نے دیکھی تھی عامر.....لیکن پتہ نہیں پھر اڑ سکی کہ نہیں؟
عامر: اڑ گئی ہوگی ارم۔پانچ دفعہ تو یہ ذکر کر چکی ہو۔اب بھول بھی تو جاؤ۔
ارم: ایسا بے وقوف ہے واجد، پتہ نہیں کس وقت جیپ میں ایئر گن چھپالی۔(فاصلے سے سائرہ اور نینی)

Figure 5: Under segmentation due to overlapping of lines

To find the line boundaries, starting at the valley index, a straight boundary is found if the image row corresponding to the valley has zero text pixels. If this is not the case, the boundary between the lines is not straight. A valley with non-zero text pixels can result due to two different situations; first when the primary ligatures across different lines touch each other and second when the primary ligatures across two lines overlap in the vertical direction (the horizontal position of two components could be different); the second case being the most common in Urdu documents. Figure 6 shows an example of touching and overlapped ligatures.


نہیں کر رہی ہو ارم۔
پتہ نہیں کیوں مجھے لگتا ہے

(a)

نہیں کر رہی ہو ارم۔
نہیں کیوں مجھے لگتا ہے

(b)

Figure 6: Problems in adjacent text lines (a): Touching ligatures (b) Overlapped ligatures

To address the issue of overlapped ligatures, we employ a non-straight traversal to segment out the lines. For each valley in the projection profile, assuming the valley position to be the center of the gap between two lines, we move in the horizontal direction searching for non-text pixels in both up and down directions. The search space is limited by the two consecutive peaks in the profile making sure that we stay within the two concerned lines. We may also come across the first scenario where there is no non-text pixel between two zones due to touching of ligatures. In such cases, we traverse the touching ligatures from the point with minimum number of text pixels. Later, the straight and non-straight line boundaries obtained from the dilated document image are superimposed on the original image to get the true text lines.

The image in Figure 5 which was under segmented due to overlapping of ligatures across multiple lines is correctly segmented using the proposed scheme as illustrated in Figure 7.

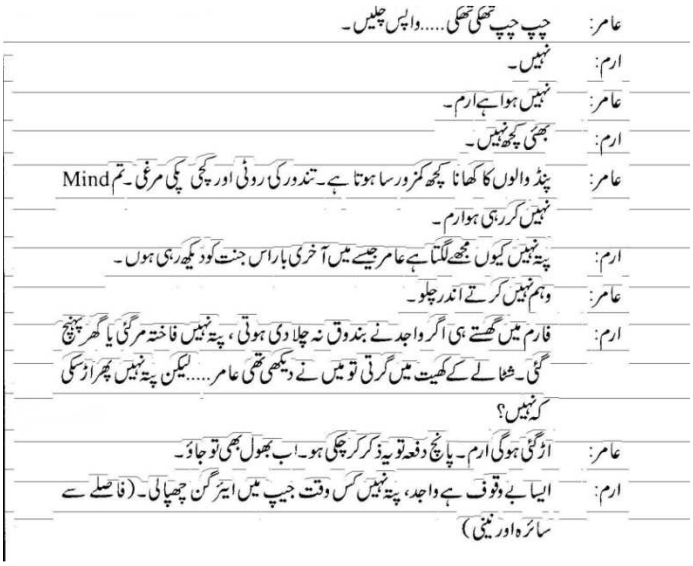

عامر: چپ چپ تھکی تھکی.....واپس چلیں۔
ارم: نہیں۔
عامر: نہیں ہوا ہے ارم۔
ارم: بھئی کچھ نہیں۔
عامر: پنڈ والوں کا کھانا کچھ کمزور سا ہوتا ہے۔تندور کی روٹی اور کچی کچی مرغی۔تم Mind نہیں کر رہی ہو ارم۔
ارم: پتہ نہیں کیوں مجھے لگتا ہے عامر جیسے میں آخری بار اس جنت کو دیکھ رہی ہوں۔
عامر: وہم نہیں کرتے اندر چلو۔
ارم: فارم میں گھستے ہی اگر واجد نے بندوق نہ چلا دی ہوتی، پتہ نہیں فاختہ مرگئی یا گھر پہنچ گئی۔شٹالے کے کھیت میں گرتی تو میں نے دیکھی تھی عامر.....لیکن پتہ نہیں پھر اڑ سکی کہ نہیں؟
عامر: اڑ گئی ہوگی ارم۔پانچ دفعہ تو یہ ذکر کر چکی ہو۔اب بھول بھی تو جاؤ۔
ارم: ایسا بے وقوف ہے واجد، پتہ نہیں کس وقت جیپ میں ایئر گن چھپالی۔(فاصلے سے سائرہ اور نینی)

Figure 7: Correct segmentation of image in Figure 5 using the proposed scheme

Sample images with touching and overlapping ligatures and their respective segmentation into lines is illustrated in Figure 8.

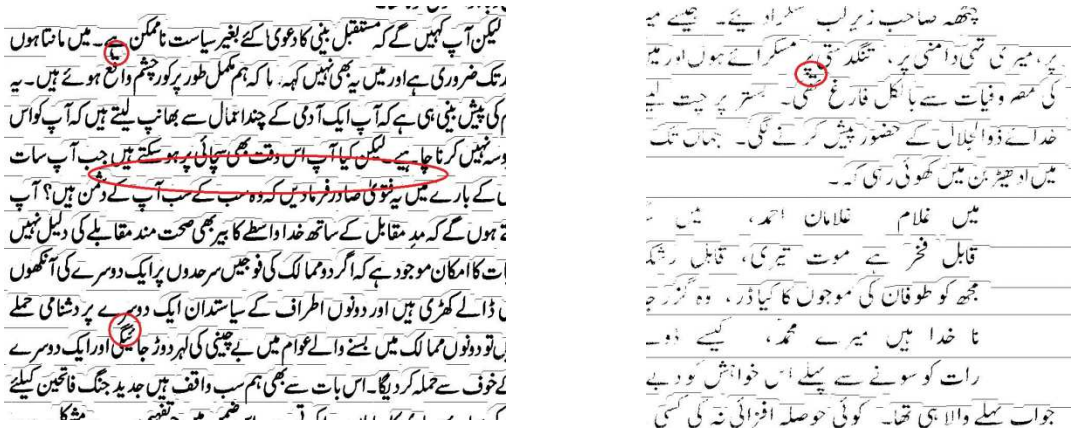

(a) (b)

Figure 8: Line segmentation (a): Contemporary image (b) Degraded image

After having discussed the line segmentation, we present the ligature segmentation in the following section.

3.2 Ligature Segmentation

Since extracting of words in Urdu text is very challenging due to non-uniform inter and intra word spacing, the recognition is typically carried out at ligature (holistic) or character (analytical approaches) levels. The segmentation into characters, however, itself is a problem in itself. Most of the Urdu OCR systems therefore use ligatures as basic unit of recognition. Traditionally, ligatures from a document are extracted by finding the connected components and applying a set of heuristics to associate the dots and diacritics with their respective primary components. These methods however, fail to provide acceptable results in case of overlapped or touching ligatures. We, therefore, propose to first segment the document into lines and then extract ligatures from each line of text separately.

Once the text lines are segmented, each line is processed to extract the primary and secondary ligatures. The major challenge in ligature extraction is the association of dots and diacritics with their respective

primary ligatures. A secondary ligature may overlap with its neighboring primary ligatures. The following steps are followed to address this issue.

- Morphological dilation is carried out on the text line with a vertical structuring element to join the dots and diacritics with their parent primary components.
- Connected components are extracted from the dilated image.
- If the dots/diacritics are joined with only one primary ligature, they are associated with that particular primary component.
- In case of an overlap with two different primary components, the horizontal distance of the secondary component(s) with the candidate primary components is computed and the secondary component is associated with the nearest primary component.

Using this technique overlapped ligatures are extracted from text-lines quite successfully an example being illustrated in Figure 9.

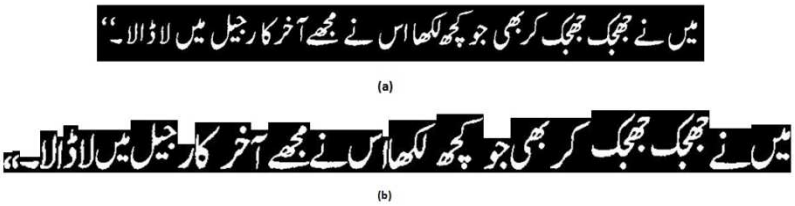


Figure 9: (a) Sample text-line (b) Extracted ligatures

4. RESULTS AND DISCUSSION

The proposed line and ligature segmentation scheme was evaluated on 30 Urdu document images from the CLE Urdu text database (www.cle.org). To evaluate the robustness of the system, 2 noisy and 3 distorted documents were included in the test set. One third of the total documents contained overlapped/touching lines. The 30 documents comprised a total of 310 text lines 306 of which were correctly segmented using the proposed technique reporting a line segmentation accuracy of 98.7%. These lines comprised a total of 7,364 ligatures out of which 6,811 were correctly segmented realizing a ligature segmentation accuracy of 92.5%.

In addition to the overall line and ligature segmentation results, we also present the results separately for each of the 30 documents as summarized in Table I. It can be seen from these results that for contemporary documents the ligature segmentation accuracy is above 90%. However, for distorted documents the accuracy drops mainly due to broken ligatures. In some cases, the ligature segmentation accuracy is low. This can be attributed to the fact that at times the secondary ligatures are not correctly associated with their respective primary ligatures leading to a fall in the segmentation accuracy.

Table I: Line and Ligature Segmentation Results on 30 Urdu Documents

S. No	Number of lines	Correctly Segmented lines	Ligatures per Document	Correctly Segmented Ligatures	Document Type	Normal/Touched/Overlapped Text lines	Accuracy
1	12	11	358	316	Contemrary	Normal	88.27%
2	7	7	207	179	Contemrary	Normal	86.47%
3	3	3	44	42	Contemrary	Normal	95.45%
4	8	8	161	156	Contemrary	Normal	96.89%
5	6	6	235	222	Contemrary	Normal	94.47%
6	10	9	96	91	Contemrary	Normal	94.79%
7	9	9	116	101	Distorted	Normal	87.07%
8	8	8	207	191	Contemrary	2 touched and 3 overlapped	92.27%
9	15	15	455	433	Contemrary	2 touched and 9 overlapped	95.16%
10	8	8	96	93	Contemrary	Normal	96.88%
11	10	10	290	270	Contemrary	4 touched	93.10%
12	17	17	460	404	Contemrary	2 touched and 7 overlapped	87.82%
13	8	8	292	284	Contemrary	Normal	97.26%
14	15	13	198	183	Contemrary	Normal	92.42%
15	8	8	262	219	Distorted	2 touched and 2 overlapped	83.59%

16	10	10	196	189	Noisy	2 touched and 2 overlapped	96.42%
17	9	9	196	187	Contemrary	Normal	95.40%
18	15	15	413	399	Noisy	2 touched	96.61%
19	11	11	271	239	Contemrary	Normal	88.19%
20	9	9	63	60	Contemrary	Normal	95.23%
21	10	10	264	253	Contemrary	Normal	95.83%
22	10	10	244	198	Contemrary	Normal	81.14%
23	10	10	288	279	Contemrary	4 overlapped	96.88%
24	18	18	279	272	Contemrary	Normal	97.49%
25	10	10	189	181	Contemrary	Normal	95.77%
26	12	12	351	319	Contemrary	Normal	90.88%
27	10	10	159	147	Contemrary	2 overlapped	92.45%
28	11	11	296	285	Contemrary	Normal	96.28%
29	7	7	279	272	Contemrary	5 overlapped	97.49%
30	14	14	399	347	Distorted	6 overlapped	86.97%
Total	**310**	**306**	**7364**	**6811**			**92.49%**

5. Conclusion and Perspectives

We presented an effective technique for segmentation of printed Urdu text into lines and ligatures, an important preprocessing step in Urdu OCR systems. The proposed technique is able to handle overlapping and touching lines and extracts lines and ligatures with high segmentation accuracies. In our further work on this problem, we intend to develop the complete ligature recognition engine where the ligatures segmented from a document image are recognized and rendered in a text editor. We also plan to evaluate the methodology on a much larger database to study the scalability of the proposed approach. It would also be interesting to expose the developed technique to the more challenging scenario of handwritten text to study how the performance varies with varying writing styles of different writers. It is expected that the segmentation methodology introduced in this study would serve to be a useful preprocessing step for researchers working on cursive OCR systems in general and Urdu OCR in particular.

REFERENCES

[1] Hande Adiguzel, Emre Sahin, and Pinar Dugulu. A hybrid approach for line segmentation in handwritten documents. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, pages 503-508.IEEE, 2012.

[2] Syed Sahib Bukhara, Faisal Safari, and Thomas M Breuer. High performance layout analysis of Arabic and Urdu document images. In Document Analysis and Recognition (ICDAR), 2011 International Conference on, pages 1275-1279. IEEE, 2011.

[3] Fazio Irbil, Aisha Latin, Nazi Knawel, and Tatyana Altar. Conversion of Urdu nastily to roman Urdu using or. In Interaction Sciences (ICIS), 2011 4th International Conference on, pages 19-22. IEEE, 2011.

[4] Sofia Tariq Javed and Swarmed Hussain. Improving nastalique speci_c pre- recognition process for urdu ocr. In Multitopic Conference, 2009. INMIC 2009. IEEE 13th International, pages 1-6. IEEE, 2009.

[5] Gurpreet Singh Lehal. Ligature segmentation for urdu ocr. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, pages 1130-1134. IEEE, 2013.

[6] Hamna Malik and Muhammad Abuzar Fahiem. Segmentation of printed urdu scripts using structural features. In Visualisation, 2009. VIZ'09. Second International Conference in, pages 191-195. IEEE, 2009.

[7] U Pal and Anirban Sarkar. Recognition of printed urdu script. In 2013 12th International Conference on Document Analysis and Recognition, volume 2, pages 1183-1183. IEEE Computer Society, 2003.

[8] Shuwair Sardar and Abdul Wahab. Optical character recognition system for urdu. In Information and Emerging Technologies (ICIET), 2010 International Conference on, pages 1-5. IEEE, 2010.

[9] Inam Shamsher, Zaheer Ahmad, Jehanzeb Khan Orakzai, and Awais Adnan. Ocr for printed urdu script using feed forward neural network. The Proceedings of World Academy of Science, Engineering and Technology, 23, 2007.

[10] Aamir Wali and Sarmad Hussain. Context sensitive shape-substitution in nastaliq writing system: Analysis and formulation. In Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, pages 53-58. Springer, 2007.